Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting

Year: 2023 | *Citations: 345* | *Authors: Zhen Qin, R. Jagerman, Kai Hui*

Abstract

Ranking documents using Large Language Models (LLMs) by directly feeding the query and candidate documents into the prompt is an interesting and practical problem. However, researchers have found it difficult to outperform fine-tuned baseline rankers on benchmark datasets. We analyze pointwise and listwise ranking prompts used by existing methods and argue that off-the-shelf LLMs do not fully understand these challenging ranking formulations. In this paper, we propose to significantly reduce the burden on LLMs by using a new technique called Pairwise Ranking Prompting (PRP). Our results are the first in the literature to achieve state-of-the-art ranking performance on standard benchmarks using moderate-sized open-sourced LLMs. On TREC-DL 2019&2020, PRP based on the Flan-UL2 model with 20B parameters performs favorably with the previous best approach in the literature, which is based on the blackbox commercial GPT-4 that has 50x (estimated) model size, while outperforming other LLM-based solutions, such as InstructGPT which has 175B parameters, by over 10% for all ranking metrics. By using the same prompt template on seven BEIR tasks, PRP outperforms supervised baselines and outperforms the blackbox commercial ChatGPT solution by 4.2% and pointwise LLM-based solutions by more than 10% on average NDCG@10. Furthermore, we propose several variants of PRP to improve efficiency and show that it is possible to achieve competitive results even with linear complexity.